文章编号:1671-4598(2014)07-2288-03　　中图分类号:TP393　　文献标识码:A

# 基于自适应权值的数据报指纹特征识别与发现

琚玉建,谢绍斌,张　薇
(空军工程大学 信息与导航学院,西安　710077)

**摘要**:面对未知协议下的报文数据,由于不能通过协议规范获得相关特征,导致传统的模式匹配方法在报文提取和协议识别过程中存在着难题;提出了以数据挖掘理论为基础的数据报指纹特征提取方案;在特征序列挖掘过程中引入自适应权值,对源数据中的序列模式进行加权统计得到判决结果;再利用提升率对特征序列进行关联规则验证,输出数据报的指纹特征;最后,采用ARP广播帧和ICMP数据包作为原始数据,测试提取数据报指纹特征;实验结果表明,自适应权值的引入能够有效减小报文中冗余数据段的干扰,提高指纹特征提取的正确率,并对报文的长度变化有一定的鲁棒性。

**关键词**:权值;自适应;未知协议;指纹特征;比特流

# Identification of Data Fingerprint Characteristics Based on Self-adaptive Weights

Ju Yujian, Xie Shaobin, Zhang Wei
(Institute of Information and Navigation, Airforce Engineering University, Xi'an　710077, China)

**Abstract**: Faced with the packet data under unknown protocol, it brought problems in the process of packet extraction and protocol identification for the traditional pattern matching method, for the reason that it couldn't obtain the relevant characteristics through protocol specification. A method for the extraction of datagram fingerprint characteristics was proposed based on data mining theory. In the process of characteristic sequence mining, it introduced the self-adaptive weights to get the verdict after the weighted statistics of sequence model from the original data. And it used Up-rate to verify the association rules between the characteristic sequence. Then fingerprint characteristics was exported. Finally, ARP broadcast frames and ICMP packets were used as raw data, and the fingerprint characteristics were extracted. Experiment results show that, the self-adaptive weights could reduce the interference of redundant data segments, improve the accuracy of the extraction of fingerprint characteristics, and have some robustness to the packet length change.

**Keywords**: weights; self-adaptive; unknown protocol; fingerprint characteristics; bit stream

## 0　引言

随着网络技术的高速发展和应用环境的多样化,网络空间的安全形势也趋于复杂。为了保障网络通信的质量和安全,对网络数据进行协议识别并提取报文进行分析是一种非常有效的途径[1]。根据特定协议的相关标准可以获得该协议下的报文特征,传统的模式匹配方法在识别分析过程中能够取得较好的效果[2]。然而当前使用的协议大多是未知协议,无法获取协议的描述文档,缺乏报文数据的相关特征,使现有的协议识别与分析手段面临着难题[3]。因此,许多研究学者在未知协议识别与分析方面进行了相关探索和研究[4-6],并取得很大进展但仍存在一定的不足,主要体现适用范围较小和易受冗余数据干扰两个方面。

为了有效减小冗余数据段的干扰,并考虑到报文的长度变化,本文提出一种基于自适应权值的数据报指纹特征提取方案。在特征序列挖掘过程中引入自适应可变的次数权值,以控制冗余数据段内的序列模式加权值,从而有效减小其对挖掘结果的干扰,同时针对报文的长度变化进行自适应调整;对挖掘到的候选特征序列以提升率[7]作为准则进行关联规则验证,进而提高指纹特征提取的正确率。仿真测试表明,该方案对于长度可变报文的指纹特征提取是有效的。

**收稿日期**:2014-03-10;　**修回日期**:2014-04-12。
**基金项目**:国家自然科学基金(61202490)。
**作者简介**:琚玉建(1990-),男,河北衡水人,硕士研究生,主要从事协议识别与分析方向的研究。

## 1　特征挖掘相关理论

### 1.1　特征序列挖掘

本文所研究的报文数据是比特流层面的,其由大量报文段无间隔地首尾相连组成。由于报文数据在发送端是在协议规范的作用下产生,因此比特序列并不是完全随机无规律可循的。在一定的监测时间内,报文头部的部分关键域(如地址域、标识域等)有较低的变化率,因而这些关键字段的序列模式成为报文数据的特征序列。运用数据挖掘中的频繁集理论,对比特流层面的报文数据进行频繁序列挖掘,即可提取协议报文的特征序列。

设比特序列 $S$ 的长度为 $l$,某序列模式 $P$ 的长度为 $m$。$S$ 中共有 $l-m+1$ 个长度为 $m$ 的序列,长度为 $m$ 的序列最多有 $2^m$ 种序列模式。则可作如下定义:

定义1:支持度。设序列模式 $P$ 在比特序列 $S$ 中出现了 $k$ 次,则序列模式 $P$ 在比特序列 $S$ 中的支持度 $Supp(P)$ 为 $\frac{k}{l-m+1}$。

定义2:频繁序列。设用户给定的支持度门限为 $\theta$,则当序列模式 $P$ 的支持度满足 $\frac{k}{l-m+1}>\theta$ 时,称序列模式 $P$ 为

频繁序列。在本文中频繁序列的支持度门限设为 $\theta = \frac{1}{2^m} \times \sigma$，其中 $\sigma$ 是支持度控制参数。

### 1.2　关联规则验证

经过比特流层面报文数据的频繁序列挖掘，可获得候选的特征序列模式。根据协议报文的特点可知，各报文段之间以及报文段内部各关键域之间存在一定的关联特性。为了有效地提取数据报的指纹特征，可以对频繁序列挖掘结果进行关联规则验证，获得更多的可靠信息。

设有序列模式 $P$ 和 $Q$，在比特序列 S 中 $P$、$Q$ 和 $P$ 与 $Q$ 先后同时出现的支持度分别为 $Supp(P)$、$Supp(Q)$ 和 $Supp(P \cup Q)$，则可作如下定义。

定义 3：置信度。关联规则 $P \Rightarrow Q$ 的置信度为 $P$ 出现的情况下，$Q$ 出现的条件概率，如公式（1）所示。

$$Confi(P \Rightarrow Q) = \frac{Supp(P \cup Q)}{Supp(P)} \tag{1}$$

由于置信度只考虑了 $P$ 出现时 $Q$ 出现的条件概率，而没有与 $P$ 不出现时 $Q$ 出现的条件概率作比较，因此无法充分地描述规则 $P \Rightarrow Q$ 的可信程度，为此引入提升率 $Up(P \Rightarrow Q)$。

定义 4：提升率。关联规则 $P \Rightarrow Q$ 的提升率为 $P$ 出现的情况下 $Q$ 出现的条件概率与 $P$ 不出现的情况下 $Q$ 出现的条件概率的比值，如公式（2）所示。

$$Up(P \Rightarrow Q) = \frac{P(Q \mid P)}{P(Q \mid \overline{P})} = \frac{Confi(P \Rightarrow Q)}{Confi(\overline{P} \Rightarrow Q)} \tag{2}$$

提升率为 1，说明 $P$ 是否出现对 $Q$ 的出现没有影响；提升率大于 1，说明 $P$ 的出现诱导了 $Q$ 的出现，关联特性较强；提升率小于 1，说明 $P$ 的出现抑制了 $Q$ 的出现，关联特性较弱。

以数据挖掘理论为基础，对报文数据的特征序列进行挖掘，并设置提升率门限对特征序列之间的关联规则进行验证，可以得到提取数据报指纹特征的重要依据。

## 2　算法设计

网络通信中的大多数协议的数据报由报文头部和报文数据段组成，如图 1 所示。在特征挖掘过程中，报文数据段（即冗余数据段）对挖掘结果造成干扰，降低了数据报指纹特征提取的可靠性。为此，设计基于自适应权值的指纹特征提取方案，如图 2 所示。

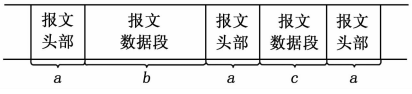


图 1　数据报通用格式

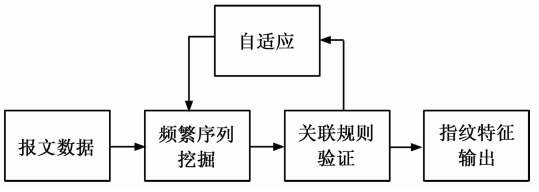


图 2　指纹特征提取流程

指纹特征提取过程主要分为 3 个模块：频繁序列挖掘模块、关联规则验证模块以及自适应模块。其中频繁序列挖掘模块包括序列模式统计算法和判决机制；关联规则验证模块包括不同频繁序列之间和相同频繁序列之间的关联规则验证；自适应模块则根据频繁序列挖掘和关联规则验证的输出结果对频繁序列挖掘阶段的相关参数进行自适应调整；最后输出数据报的指纹特征。

### 2.1　基于散列的序列模式统计

对比特序列的统计过程中面临着大数据量中查询匹配序列模式的时空复杂度较高的问题，单模式匹配方法必须枚举所有序列模式逐一扫描匹配。为此设计基于散列的序列模式统计算法，将序列模式按位长分组进行统计，并以序列模式转换为的十进制整型数字 $i$ 为关键字散列到数组中进行存储。以长度为 $m$ 的序列模式组统计为例，算法描述如下：

输入：源报文数据的比特序列 $S$、序列模式长度范围 min～max、权值 $\varepsilon_0$；

输出：各序列模式加权值 $P_\varepsilon(i)$、序列出现位置 $pos$；

流程：

（1）定义统计结构体，结构体中包含序列模式、关键字、加权值、位置信息以及指针等内容。

（2）枚举长度为 $m$ 的序列模式，散列到结构数组中以备查询统计。

（3）遍历比特序列 $S$ 中长度为 $m$ 的序列模式，更新结构数组中对应元素状态。

（4）遍历结束，输出结果。

统计算法的整个流程可用图 3 描述。

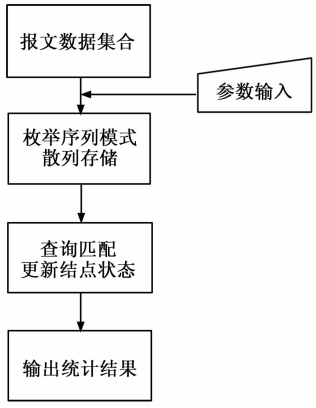


图 3　序列统计算法流程图

通过基于散列的序列模式统计过程，可以仅对源报文数据扫描 max～min 次即可获得所有序列模式的加权统计结果，并与给定的支持度门限 $\theta$ 比较即可获得频繁序列模式。实现过程中，为节约内存消耗，当 $m$ 较大时可使用除法散列函数进行再散列，如公式（3），并用链式散列解决冲突问题。

$$remainder = \frac{i}{2^{20}} (m > 20) \tag{3}$$

### 2.2　提升率计算

通过频繁序列挖掘模块可以获得候选的特征序列模式，但由于数据段的干扰以及人为参数选择的原因，难免存在一些序列模式对于指纹特征提取并不是有意义的。因此，需要进一步验证序列模式间的关联规则。本文采用提升率来衡量序列模式间的关联特性。

假设某项关联规则 $P \Rightarrow Q$ 的含义为序列模式 $P$ 出现后的 64 位内序列模式 $Q$ 出现，并由 $P(Q \mid P)$ 表征此条件概率。而

$P(Q \mid \overline{P})$ 表征序列模式 $Q$ 出现之前的 64 位内序列模式 $P$ 并没有出现的条件概率。则通过公式（2）即可计算关联规则 $P \Rightarrow Q$ 的提升率 $Up(P \Rightarrow Q)$，当其大于给定的提升率门限时，则认为这条关联规则是有意义的。对不同序列模式 $P$ 和 $Q$ 进行关联规则验证，可以获得更多的报文数据指纹特征，提高推断报文头部位置的正确率。对同一序列模式 $P_j(j = 1,2,3,\ldots)$ 进行关联规则验证，根据频繁序列集中出现的位置合理指定位置差考察范围，则可推断报文的长度 $N$。

以给定的提升率门限为准则对频繁序列模块挖掘到的序列模式进行关联规则验证，可以进一步获取数据报的指纹特征信息，并可输出报文的可能长度 $N$。

### 2.3 基于自适应权值的指纹特征提取

为了更有效地降低冗余数据段对指纹特征提取的干扰，并考虑到报文的长度变化，设计自适应可变的权值 $\varepsilon$。在自适应模块中，引入以下参数。

粒度参数 $G$：控制权值 $\varepsilon$ 不同取值的位长范围；

反馈周期 $N_T$：控制自适应周期，本文取 $N_T = 50 \times N$，$N$ 为关联规则验证模块输出的报文长度；

序列密度 $\rho_F$：表征粒度 $G$ 范围内频繁序列模式出现的密度，设出现频繁序列模式个数为 $F$，即有 $\rho_F = \dfrac{F}{G}$。分析可知，$\rho_F \in (0, \max \sim \min)$。

自适应过程可描述为：频繁序列挖掘模块获得候选的特征序列以及出现的位置；再由关联规则验证模块对报文长度进行推断；根据粒度参数将每个报文分块计算序列密度，建立序列密度 $\rho_F$ 到权值集合 $\varepsilon = \{0.2, 0.4, 0.6, 0.8, 1\}$ 的映射；设置反馈周期，即可根据自变量 $\rho_F$ 的大小，输出不同的函数 $\varepsilon$ 值作用于频繁序列挖掘模块，完成自适应过程。自适应过程的整个流程可用图 4 描述。

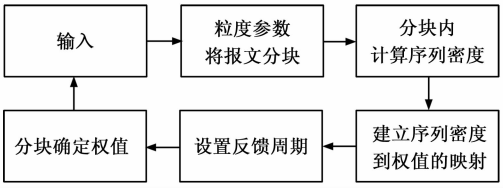


图 4 自适应模块流程图

在指纹特征提取流程中，输入源报文数据和序列模式长度范围并对相关参数进行设定，即可输出报文数据的指纹特征，包括：特征序列模式、帧长以及序列模式间有意义的关联规则。

## 3 仿真测试

为了验证其有效性，算法采用 VC＋＋6.0 编程实现，测试数据由 Wireshark 工具在局域网环境内抓包获取。本文采用由 ARP 广播帧和 ICMP 数据包组成的报文数据集合进行测试，数量均为 200 个。

### 3.1 筛选率与命中率

通过 Wireshark 对报文数据集合的解析结果可以对频繁序列进行预测，并得到预测频繁序列集合。定义筛选率 $\upsilon$ 为剔除的序列模式与总序列模式个数的比值，命中率 $\mu$ 为提取的预测频繁序列在频繁序列集中所占的比例。对两种不同报文组成的数据集合提取 8 位频繁序列，支持度参数取值范围为 0.1～2.0，针对无权值、权值 $\varepsilon = 0.6$、权值自适应调整三种情况分别进行测试。筛选率曲线如图 5 所示，命中率曲线如图 6 所示。

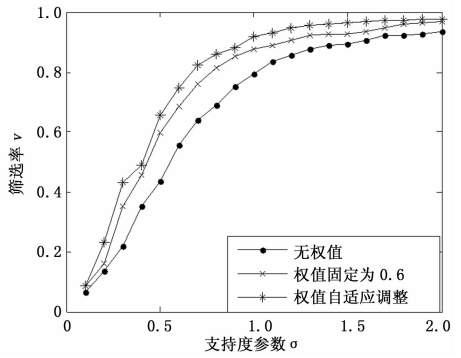


图 5 筛选率曲线

分析图 5 可知，随着支持度参数的增大，筛选率在 3 种情况下均有提高，且权值的引入优于无权值的情况，权值自适应调整优于权值固定的情况，充分说明了自适应权值的引入使挖掘结果对于提取指纹特征更有参考价值；此外，当 $\sigma < 0.8$ 时，筛选率提升较快，而当 $\sigma > 0.8$ 时，筛选率变化缓慢，这是因为当支持度门限提升至一定范围时，大部分非频繁序列模式被剔除，而只保留了频繁序列模式。

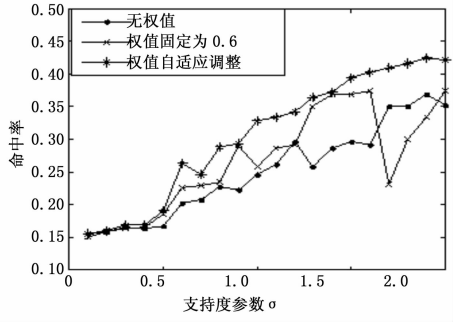


图 6 命中率曲线

分析图 6 可知，随着支持度参数的增大，命中率在 3 种情况下均曲折变化，总体趋势均有提高，且权值的引入优于无权值的情况，权值自适应调整优于权值固定的情况，充分说明了自适应权值的引入使得算法对于冗余数据段的干扰控制更为有效；此外，当 $\sigma$ 在 0.7～1.2 范围内，权值固定与无权值的情况上下浮动较大，而权值自适应调整的情况浮动较小，浮动原因是支持度提高至一定范围时，频繁序列集合中会丢失部分预测频繁序列，从而使命中率上下浮动。再次说明权值自适应调整较权值固定更适合长度可变报文的指纹特征提取。

### 3.2 报文长度推断

为了提取更多的数据报指纹特征，可进一步对挖掘到的特征序列进行关联规则验证。对报文数据集测试提取 20～24 位频繁序列，支持度参数设置为 0.8，对挖掘得到的频繁序列集合进行相同频繁序列间的关联规则验证，提升率门限设置为 1.05。绘制不同位置差值 $d$ 的出现次数 $f$，推断报文长度范围如图 7 所示。

（下转第 2294 页）

是能到达 LS 同步方法的 5 倍以上。得出结论，在低信噪比环境下整体最小二乘法比普通最小二乘法更合适于做同步误差估计。

**参考文献：**


[1] Vannee D J R，Coenen A J R M. New fast GPS code－acquisition technique using FFT [J]. Electronics Letters，1991，27：158－160.

[2] Kaplan E D. Understanding GPS：Principles and Application [M]. Boston：Artech House Publishers，1996.

[3] 龚国辉，李思昆. 提高 DSSS 信号 PN 码相位测量精度的三点二次插值法 [J]. 通信学报，2007，28 (2)：130－133.

[4] 胡修林，曾 臻，张 俊，等. 直扩系统伪码精确同步及 FPGA 实现 [J]. 华中科技大学学报（自然科学版），2005，33 (6)：44－46.

[5] Golub G H. Van Loan C F. An analysis of the Total Least Squares problem [J]. SIAM J Numer Anal，1980，17 (6)：883－893.

[6] Van Huffe Sl，Vandewalle J. The Total Least Squares Problem Computational Aspects and Analysis [M]. SIAM，Philadelphia，1991.

[7] 丁克良，沈云中，欧吉坤. 整体最小二乘法直线拟合 [J]. 辽宁工程技术大学学报（自然科学版），2010，29 (1)：44－47.

[8] 李红伟，魏少春. 陈安平，等. 总体最小二乘法在直线拟合中的应用 [J]. 地矿测绘，2010，26 (2)：4－5.

[9] 王徐华，柏 鹏，李寰宇，等. 基于多速率内插和最小二乘法的精确同步方法 [J]. 计算机应用研究，2012，29 (10)：3922－3925.


（上接第 2290 页）

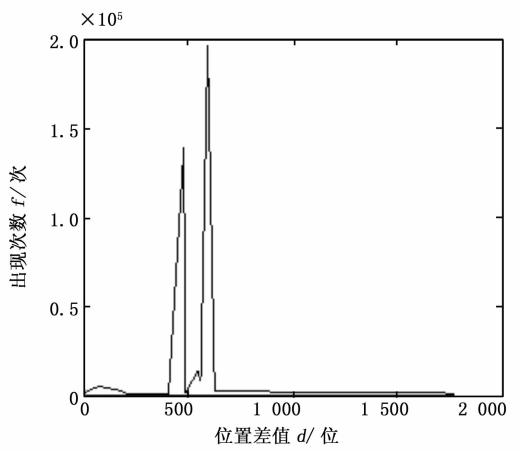


图 7 报文长度推断

分析图 7 可知，位置差值 592 与位置差值 480 出现次数远多于其他差值，且前者多于后者。分析 Wireshark 对两种报文的解析结果可知，报文长度推断正确，进一步验证了算法的有效性；其中位置差值 592 出现次数更多的原因是 ICMP 报文的特征序列位长大于 ARP 报文的特征序列，而提取的频繁序列位长较短，使得同一段特征序列重复计数。

## 4 小结

本文以数据挖掘中频繁集和关联规则相关理论为基础，针对长度可变的报文集合设计了基于自适应权值的数据报指纹特征提取方案。真实数据测试表明，该方案对于数据报的指纹特征提取是有效的，自适应权值的引入使得筛选率和命中率都有提高且更为稳定，有效控制了冗余数据段对于提取结果的干扰。该方案在实际应用中还应该根据实际情况合理设置参数以有效可靠地提取指纹特征[8-14]。

**参考文献：**


[1] Kim M S，Won Y J，Hong J W K. Application－level traffic monitoring and an analysis on IP networks [J]. ETRIjournal，2005，27 (1)：22－42.

[2] Callado A，Kamienski C，Szabo G，etal. A survey on internet traffic identification [J]. Communications Surveys & Tutorials，IEEE，2009，11 (3)：37－52.

[3] Trifilo A，Burschka S，Biersack E. Traffic to protocol reverse engineering [A]. Computational Intelligence for Security and Defense Applications，CISDA 2009，IEEE Symposium on. IEEE [C]，2009：1－8.

[4] 张一嘉. 局域网链路层数据帧识别算法的设计与实现 [J]. 通信对抗，2007，(4)：41－44.

[5] 白 彧，杨晓静，张 玉. 基于高阶统计处理技术的 m－序列帧同步码识别 [J]. 电子与信息学报，2012，34 (1)：33－37.

[6] 金 凌. 面向比特流的未知帧头识别技术研究 [D]. 上海：上海交通大学，2011.

[7] 马占欣，王新社，黄维通，等. 对最小置信度门限的置疑 [J]. 计算机科学，2007，34 (6)：216－218.

[8] 王祥斌. 数据挖掘技术在入侵检测系统中的应用研究 [J]. 计算机测量与控制，2012，20 (2)：321－323.

[9] 张文博，姬红兵，王 磊. 一种自适应权值的多特征融合分类方法 [J]. 系统工程与电子技术，2013，35 (6)：1133－1137.

[10] 裴颂文，吴百锋. 动态自适应特征权重的多类文本分类算法研究 [J]. 计算机应用研究，2011，28 (11)：4092－4096.

[11] Jiang H C，Xu J Z，Shi W S，et al. Stellar spectra association rule mining method based on the weighted frequent pattern tree [J]. Research in Astron. Astrophys，2013，13 (3)：3434－342.

[12] 张昆明，甘文丽，李元臣. Master－Worker 模式的并行关联规则挖掘算法 [J]. 计算机测量与控制，2013，21 (4)：1008－1010.

[13] Tian X G，Duan M Y，Sun C L，et al. Detecting network intrusions by data mining and variable－length sequence pattern matching [J]. Journal of SystemsEngineering and Electronics，2009，20 (2)：405－411.

[14] 刘兴彬，杨建华，谢高岗，等. 基于 Apriori 算法的流量识别特征自动提取方法 [J]. 通信学报，2008，29 (12)：51－59.